数据采集规则报告

**摘要**

在全球化加速的背景下，中国的经济和科技力量正在迅速崛起，主要经济体之间的 竞争愈发激烈，尤其是英美日法等国家在对华政策上的立场备受关注。为了深入分析这 些国家对华政策的变化及其可能产生的影响，我们选择了 2023-2024 年美国在经济、文 化、立法和科技等领域对华态度作为研究主题。本报告将介绍数据采集方案，包含关键 词筛选、微博内容的优化筛选、数据存储与处理等关键环节，为相关领域的深入研究提 供数据支持和理论基础。

**1 采集分析**

**1.1 介绍**

在全球化加速的今天，中国的经济和科技实力正日益崛起，与此同时，主要经济体 之间的竞争也日趋激烈。尤其是英美日法等国家在对华政策上的态度备受瞩目。为了深 入研究这些国家对华政策的变化趋势及其潜在影响，我们选择以 2023-2024 年美国在经 济、文化、立法和科技等领域的对华态度为研究主题。

这一研究不仅具有重要的学术意义，还能为政策制定提供宝贵的参考依据。通过分 析美国在经济制裁、技术封锁、文化交流、立法监管等方面的动态数据，我们可以全面 了解当前的国际竞争格局，并为中国制定相应的应对策略提供数据支持和理论基础。

**1.2 数据的可采集性分析**

在研究过程中， 数据的获取和代表性至关重要。由于时间和数据来源的限制，我们 主要聚焦于社交媒体和官方新闻网站的数据。这些数据具有较高的参考价值。我们一共 采取了 80000 多条数据。

• **微博**

**–** 实时性强：微博数据能够及时反映社会热点和政策变化。

**–** 数据量大且多样化：微博平台拥有庞大的用户基础，数据类型多样，包括短 文、评论、图片和视频等。

**–** 非结构化数据多：大量非结构化数据为自然语言处理提供了理想的场景，能 从中提取出有价值的信息和观点。

• **官方新闻网站**

**–** 数据质量高：新闻网站内容经过严格审查，具有权威性、真实性，适合进行 定量和定性分析。

**–** 公信力高：选取如人民网、新华网等权威新闻平台，确保数据的可靠性。

**–** 内容规范：这些平台遵守新闻报道规范，避免了主观倾向，内容适合数据分 析。

**2 数据采集规则**

**2.1 介绍**

我们将基于 Scrapy 框架设计一个爬虫程序， 以实现对微博平台上指定关键词的搜 索结果进行数据采集。目标是获取包含指定关键词且在指定时间范围内（2023 年-2024 年）发布的微博信息，并将其保存到本地文件。同时，我们还将使用八爪鱼软件进行数 据爬取，由于其操作简单且具备可视化功能，能够有效辅助我们的数据采集工作。

**2.2 工具**

为了实现高效的数据采集，我们使用以下工具：

• **Python 爬虫**：采用 Scrapy 框架，主要特点是爬取速度相对较快，支持高并发请 求。

• **八爪鱼软件**：操作简单方便，适合非技术背景用户进行数据抓取。

**2.3 关键词选择及筛选策略**

• **领域细分：**扩展更多与分析主题相关的细分关键词，例如“中美科技竞争”、“美 国对华立法政策”、“中美文化”等，确保全面覆盖主要研究领域。

• **跨领域整合：**整合不同领域的关键词，确保涵盖与分析主题相关的多种话题，如 经济政策、外交关系、文化交流等，以提高数据的多样性和丰富性。

**2.4 微博类型与内容筛选优化**

• **原创微博优先：** 优先采集原创微博，避免因转发内容重复导致数据冗余.

• **内容类型多样化：**在原有的文本信息采集基础上，同时抓取包含图片和视频的微 博内容，以提供更丰富的分析素材，这为未来进一步的多模态领域的分析做了准 备。

**2.5 反爬虫策略优化**

• **代理池管理：**构建动态代理池，自动切换不同 IP 地址，避免因频繁请求被平台封 禁。

• **请求延迟与限速：**引入随机延迟和限速策略，避免短时间内发送过多请求而触发 反爬机制。

• **随机用户代理：**模拟多种浏览器的请求头，增强爬虫的随机性，减少被平台识别 为爬虫的风险。

**2.6 爬取目标**

我们的数据采集主要集中在以下两个方面：

**2.6.1 微博**

关键词：我们将搜索在微博正文中包含的指定关键词，例如：“中美科技”、“中美 政治”、“中美文化”、“中美经济”、“中美国际局势”、“美国对华经济政策”、“美国对华 科技政策”、“美国对华文化政策”、“美国对华立法政策”、“中美关系”、“中国美国”。 时间范围为 2023 年 1 月 1 日至 2024 年 10 月 5 日。

**2.6.2 人民网等其他网站**

关键词：在新闻标题和正文中包含指定关键词，具体关键词为：“中美科技”、“中 美政治”、“中美文化”、“中美经济”、“中美国际局势”、“美国对华经济政策”、“美国对 华政策”、“中美关系”、“中美外交”。时间范围同样为 2023 年 1 月 1 日至 2024 年 10 月 5 日。

**2.7 采集内容**

我们将对以下内容进行采集：

**2.7.1 微博**

• 微博 ID：唯一标识微博的一串数字形式。

• 微博 bid：微博的简短标识符。

• 用户昵称：微博发布者的名称。

• 微博正文：微博正文文本。

• 话题：微博正文中提到的话题内容。

• 转发数：微博被转发的次数。

• 评论数：微博的评论数量。

• 点赞数：微博获得的点赞数量。

• 发布时间：微博的具体发布时间。

• 微博图片 URL：微博的图片 URL 地址。

• 微博视频 URL：微博的视频 URL 地址。

• 发布位置：用户发布微博时的地理位置信息。

**2.7.2 新闻网、人民网等**

• 标题：新闻的标题。

• 发布时间：新闻发布的年份。

• 正文：新闻的主体内容。

• 来源：发布新闻报道的官方。

**2.8 保存方式**

数据将被保存为以下格式：

• 保存微博数据到.xlsx 文件中，以便于后续数据分析。

• 保存新闻数据到.csv 文件中，以便于数据处理。

**2.9 使用技术**

为了高效地实现数据抓取，我们将采用以下技术：

• 框架：使用 Python 的 Scrapy 框架来实现爬虫逻辑， 支持高效的数据抓取和并发 请求。

• 数据解析：通过 XPath 或 CSS 选择器进行数据解析。

• 用户代理轮换：为避免被封禁，使用多种用户代理进行请求头设置。

• 反爬策略：引入延迟和代理池，以减少爬虫被检测到的风险。

**2.10 Scrapy 项目框架**

Scrapy 项目的标准目录结构如下：

weibo\_scraper

+-- weibo\_scraper/

| +-- \_\_init\_\_.py

| +-- items.py

| +-- middlewares.py

| +-- pipelines.py

| +-- settings.py

| +-- spiders/

| +-- weibo\_spider.py

+-- scrapy.cfg

• items.py：定义要抓取的数据结构。

• pipelines.py：处理数据保存的逻辑，包括 CSV 文件。

• settings.py：包含爬虫的全局配置，如用户代理、下载延迟、并发数等。

• spiders/weibospider.py：爬虫的主逻辑， 包括抓取网页、解析数据、处理反爬 等。

通过上述设计，爬虫将根据关键词和时间范围，在微博中搜索相关内容并提取所需 信息，最终保存到 CSV 文件。使用 Scrapy 框架，确保爬虫能够高效并发执行，并对大 规模数据进行采集和处理。

**3 数据清洗与预处理**

在数据采集后，我们对采集到的数据进行了清洗和预处理，以提升数据的质量：

• **缺失值处理**：删除微博正文缺失的数据， 保留其他字段缺失但不影响分析的记录。

• **重复数据处理**：去除重复采集的数据。

• **异常值处理**：对异常字符进行编码转换，确保内容为 UTF-8 格式。

• **数据整合**：将微博数据与新闻数据进行整合，形成统一的数据库表结构。

**4 数据可靠性与科学性**

为了确保数据的可靠性，我们从多个角度对数据的可信度进行了验证：

• **数据来源可靠**：通过 Python 爬虫和八爪鱼采集工具，从官方新闻平台和权威社 交平台（如微博）采集数据，确保数据来源的权威性。

• **处理过程科学性** ：数据经过清洗、转换和整合，处理过程科学严谨。

• **数据规模和代表性**：共采集了超过 80,000 条微博与新闻数据，覆盖广泛。

• **结果可重复性**：在不同设备和不同时间段内重复运行爬虫，数据采集过程一致。